



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

### Fish or Fowl: A Wizard of Oz Evaluation of Dialogue Strategies in the Restaurant Domain

**Citation for published version:**

Whittaker, S, Walker, M & Moore, J 2002, Fish or Fowl: A Wizard of Oz Evaluation of Dialogue Strategies in the Restaurant Domain. in *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain..* European Language Resources Association (ELRA).

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Peer reviewed version

**Published In:**

Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain.

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Fish or Fowl: A Wizard of Oz Evaluation of Dialogue Strategies in the Restaurant Domain

Steve Whittaker\*, Marilyn Walker\*, Johanna Moore†

\* AT&T Labs - Research  
Florham Park, NJ, USA, 07932  
steve.walker@research.att.com

† University of Edinburgh  
2 Buccleuch Place  
Edinburgh, Scotland, EH8 9LW  
jmoore@cogsci.ed.ac.uk

## Abstract

Recent work on evaluation of spoken dialogue systems suggests that the information presentation phase of complex dialogues is often the primary contributor to dialogue duration. This indicates that better algorithms are needed for the presentation of complex information in speech. Currently however we lack data about the tasks and dialogue strategies on which to base such algorithms. In this paper, we describe a Wizard of Oz tool and a study which applies user models based on multi-attribute decision theory to the problem of generating tailored and concise system responses for a spoken dialogue system. The resulting Wizard corpus will be distributed by the LDC as part of our work on the ISLE project.

## 1. Introduction

Recent work on evaluating spoken dialogue systems suggests that the information presentation phase of complex dialogues is often the primary contributor to dialogue duration (Walker et al., 2001). During this phase the system returns from a database query with a set of options that match the user's constraints (See Figure 1). The user then navigates through these options and refines them by offering new constraints. Depending on the number of options returned, and the facilities provided for refinement, this process may be quite onerous. Even if the dialogue system has the ability to display some information graphically, the information presentation phase can still be quite tedious because the user has to browse a list or graphical representation of the options to access information about each one. Thus dialogue systems have a critical need for better algorithms for: (1) selecting the most relevant subset of options to mention, and (2) choosing what to say about them.

One problem in developing such algorithms for complex information gathering dialogues is that we lack detailed information about (a) the types of tasks and the strategies employed by users when seeking information and (b) the techniques employed by human experts in providing information about those complex domains.

The current study addresses this problem by first col-

SYS: I found 9 round trips from Hartford to Orlando on October 31st and returning Hartford on November 6th. For option 1, I have an outbound flight with U S Air departing at 7 AM and arriving at 9 49 AM non-stop. A return flight is with U S Air departing at 10 30 AM and arriving at 12 28 AM with 1 stopover. Ticket price is 165 dollars. Please say next option or flight details or i'll take it.

USER: NEXT OPTION

SYS: For option 2, I have an outbound flight with U S Air departing at 3 30 PM and arriving at 6 19 PM non-stop. A return flight is with U S Air departing at 8 25 PM and arriving at 12 28 AM with 1 stopover. Ticket price is 165 dollars. Say next option, or, flight details, or, I'll take it.

USER: NEXT OPTION

.....

Dialogue continues until user selects option

Figure 1: Information presentation phase of a Communicator dialogue

lecting representative user tasks in the restaurant domain, and then collecting sample dialogues using the Wizard of Oz technique. We develop a novel Wizard of Oz tool to assist with collecting sample data in this domain.

A second goal of the study is to investigate the utility

of employing a user model based on multi-attribute decision theory (Keeney and Raiffa, 1976; Edwards and Barron, 1994), and dialogue strategies based on the user model when providing information in this domain (Carenini and Moore, 2000; Carenini and Moore, 2001). In two dialogues, the Wizard dialogue exploits the user model to generate first a user-tailored SUMMARY for presenting multiple restaurant options, followed by a user-tailored RECOMMENDATION. We compare performance under those conditions with performance on two dialogues where the Wizard utilizes a default user model and a SERIAL presentation based on those used by the AT&T Communicator system in the travel planning domain (Levin et al., 2000), as illustrated in Figure 1.

The structure of the paper is as follows. Section 2. describes the development and collection of user models based on multi-attribute decision theory. We also talk about the elicitation procedure for acquiring these models. Section 3. provides details about the Wizard interface tool that was employed in the experiment to support data collection. Section 4. describes: (a) the experimental design; (b) and the tasks used and the data collected. Section 5. describes our experimental results. The study was designed to achieve the following: (a) collect representative tasks for a complex information seeking domain; (b) test the utility of a combination of a user model/dialogue strategy in that domain. The dialogues we collected will be made available as part of the ISLE program.

## 2. Multi-Attribute Decision Models in the Restaurant Domain

Multi-attribute decision models are based on the fundamental claim that if anything is valued it is valued for more than one reason (Keeney and Raiffa, 1976). In the restaurant domain, this implies that a user's preferred restaurants optimize a combination of restaurant attributes. In order to define a multi-attribute decision model for the restaurant domain, we must determine the attributes and their relative value for particular users. Edwards and Barron describe a procedure called SMARTER for eliciting multi-attribute decision models for particular users or user groups (Edwards and Barron, 1994). First, the important attributes in the domain, and their relationships to each other, are identified. Second, the values of each attribute are mapped to single-dimension cardinal utilities that span the whole range from 0 to 100. Third, a function is defined that combines the utilities for each attribute into an overall utility score for an option. Finally, weights (or rankings) are assigned to each attribute that indicate the importance of that attribute to each user. The SMARTER procedure also specifies how to elicit these weights from users in a way that

takes little time and has been shown to result in more accurate user models than simple ranking (Edwards and Barron, 1994).

The attributes in the restaurant domain were mapped to cardinal utilities by a simple linear transformation. For categorical attributes such as food type, values that the user likes get mapped to 90, dislike values to 10 and other values to 50. The function we use to combine the utilities is a simple additive model; the value for each attribute is multiplied by its weight and all the weighted values are added up.

To elicit the user models, we carried out a telephone interaction to get the user to rank order the attributes in the domain. The user was first asked: *Imagine that for whatever reason you've had the horrible luck to have to eat at the worst possible restaurant in the city. The price is 100 dollars per head, you don't like the type of food they have, you don't like the neighborhood, the food itself is terrible, the decor is ghastly, and it has terrible service. Now imagine that a good fairy comes along who will grant you one wish, and you can use that wish to improve this restaurant to the best there is, but along only one of the following dimensions. What dimension would you choose? Food Quality, Service, Decor, Cost, Neighborhood, or Food Type?* The user would choose an attribute and the scenario would be repeated omitting the chosen attribute. The procedure was continued until all attributes had been selected, and took less than 5 minutes overall.

We elicited models in this way for 15 users; the models are stored in a database that is accessed by the Wizard program. Five examples of different user models are in Figure 2. The columns show the weightings associated with continuous variables and particular likes/dislikes for categorical variables. Note that for all five users, food quality is important, being the highest or second highest ranked attribute. Cost is also relatively important for each of these users, with both decor and service being of lesser importance. Overall in the 15 user models, food quality and cost were generally among the top three ranked attributes, while the ranking of other attributes such as decor, service, neighbourhood and foodtype varied widely.

The user model reflects a user's *dispositional* biases about restaurant selection. These can be overridden by *situational* constraints specified in a user query. For example, as Figure 2 shows, some users express strong preferences for particular food types. However these preferences can be overridden in any particular situation by simply requesting a different food type. Thus *dispositional* biases never eliminate options from the set of options returned by the database, they simply affect the *ranking* of options.

| User | Food Quality | Service | Decor | Cost | Nbhd | FT   | Nbhd Likes  | Nbhd Dislikes  | FT Likes   | FT Dislikes                                       |
|------|--------------|---------|-------|------|------|------|---|--|--|---|
| CK   | 0.41         | 0.10    | 0.03  | 0.16 | 0.06 | 0.24 | Midtown, Chinatown, TriBeCa                                 | Harlem, Bronx  | Indian, Mexican, Chinese, Japanese, Seafood        | Vegetarian, Vietnamese, Korean, Hungarian, German |
| HA   | 0.41         | 0.10    | 0.03  | 0.16 | 0.06 | 0.24 | Upper West Side, Chelsea, Chinatown, East Village, TriBeCa  | Bronx, Uptown, Harlem, Upper East Side, Lower Manhattan                    | Indian, Mexican, Chinese, Japanese, Thai           | no-dislike  |
| OR   | 0.24         | 0.06    | 0.16  | 0.41 | 0.10 | 0.03 | West Village, Chelsea, Chinatown, TriBeCa, East Village     | Upper East Side, Upper West Side, Uptown, Bronx, Lower Manhattan           | French, Japanese, Portuguese, Thai, Middle Eastern | no-dislike  |
| MSH  | 0.41         | 0.10    | 0.06  | 0.24 | 0.16 | 0.03 | Flatiron, Chelsea, West Village, Midtown East, Midtown West | Chinatown, Lower East Side, East Village, Upper East Side, Upper West Side | Indian, Mexican, Ethiopian, Thai, French           | Steakhouse, Russian, Korean, Filipino, Diner      |
| SD   | 0.41         | 0.10    | 0.03  | 0.16 | 0.06 | 0.24 | Chelsea, East Village, TriBeCa                              | Harlem, Bronx  | Seafood, Belgian, Japanese                         | Pizza, Vietnamese                                 |

Figure 2: Sample User Models: Nbhd = Neighborhood; FT = Food Type

### 3. Wizard Interaction Tool

The Wizard's task of generating dialogue responses in real time based on a specified set of strategies and a user model is quite demanding. We therefore built a Wizard interface based on XML/XSLT, to aid in data collection. We used the MATCH system (Johnston et al., 2001) to provide a backend database of New York restaurants. The interface allows the Wizard to specify a set of restaurant selection criteria, and returns a list of options that match the users' request. The tool includes the appropriate user model, allowing the Wizard to identify restaurant options and attributes that are important to the particular user. The Wizard uses this information, along with a schema of dialogue strategies to guide his interaction with the user in each dialogue. The RECOMMEND strategy that the Wizard uses in the dialogues is motivated by user tailored strategies for the real estate domain described in (Carenini and Moore, 2001).

Figure 3 illustrates the Wizard interaction tool (WIT). The main function of the interface is to provide relevant information to allow the Wizard to quickly identify sets of restaurants satisfying the user's query, along with reasons for choosing them, while respecting the particular preferences of that specific user. A major constraint on the design of WIT was to allow the users to modify an original task specification during the course of the dialogue in response to information provided by the Wizard.

The tool (see Figure 3) contains three main panels. The right hand panel supports query specification, allowing the Wizard to specify constraints corresponding to the user's query, which in this example is *Japanese, Korean, Malaysian or Thai restaurants, costing between 30-40 dollars, with food quality greater than 20 and service and decor greater than 15 anywhere in Manhattan*. The right

hand panel contains radio buttons allowing the Wizard to specify: cost range (using one button for upper, and one for lower limits), food quality, service, decor, cuisine and neighbourhood. Omitting a selection (e.g. neighbourhood) means that this attribute is unconstrained, corresponding in this case to the statement *anywhere in Manhattan*.

The left hand panel shows the restaurants satisfying the query along with information for each restaurant including its overall utility, and the absolute values for food quality, service, decor and cost. For each attribute we also supply corresponding weighted attribute values, shown in brackets after each absolute attribute value. The overall utility and weighted attributes are all specific to a given user model. The user model here is the CK model from Figure 2. In this example, for the restaurant Taka, the overall utility is 75, absolute food quality is 25 (weighted value 34), service is 23 (absolute) and 6 (weighted), decor is 15 (absolute) and 1 (weighted) and cost is 37 dollars (absolute) and 10 (weighted). So the main reason why CK should like Taka, according to the user model, is that the food quality is excellent, as indicated by the fact that this attribute contributes almost half of the overall weighted utility (34 out of 75 units).

The centre panel of Figure 3 shows specific information about the restaurant selected in the left hand panel, including its address, neighbourhood, a review and telephone number.

Overall the tool provides a method for the Wizard to quickly identify candidate restaurants satisfying a particular user's preferences, along with reasons (the weighted attribute values) why the user should choose that restaurant. The UI also allows the Wizard to see at a glance the trade-offs between the different restaurants, by comparing their

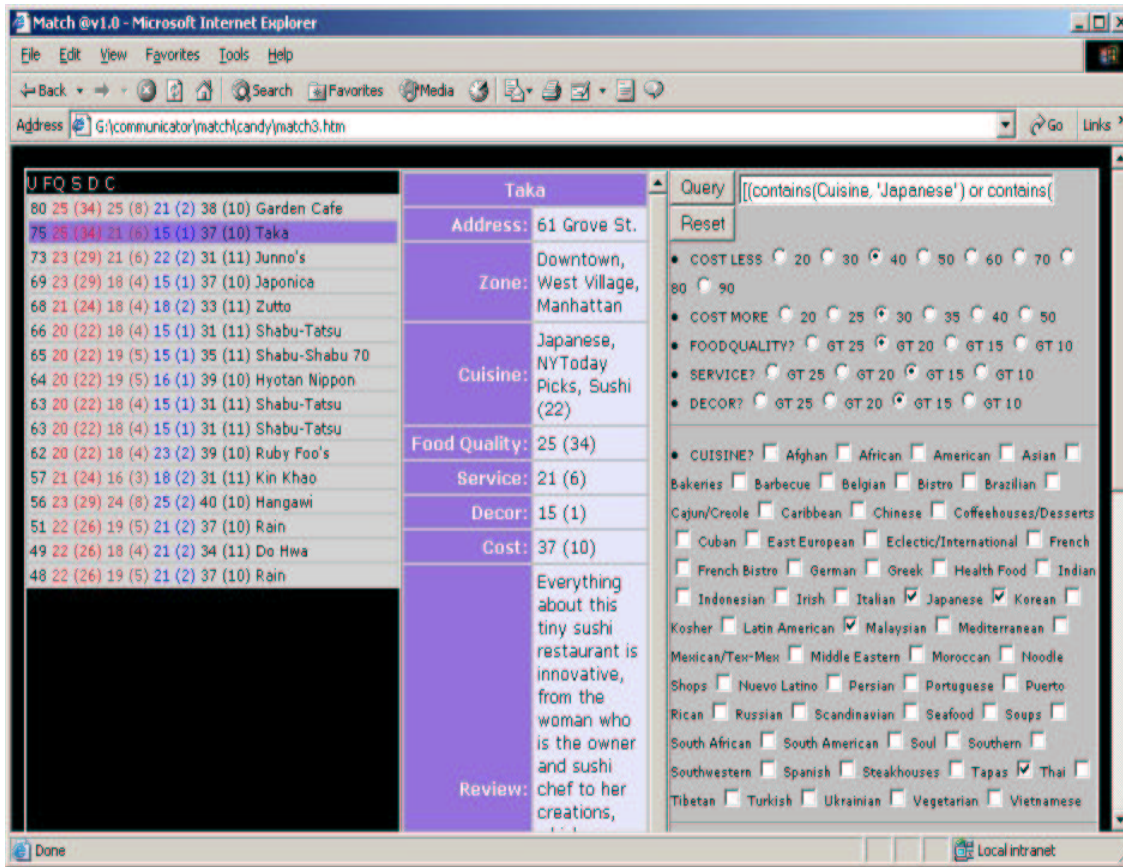


Figure 3: Wizard Interface for user CK after Wizard enters query for: *Japanese, Korean, Malaysian or Thai restaurants, costing between 30-40 dollars, with food quality greater than 20 and service and decor greater than 15 anywhere in Manhattan.*

different weighted utilities. For example, the main reason for preferring the Garden Cafe over Taka is that it has better service and decor (as shown by the different weighted values of these attributes).

We demonstrate the effects of the user model by showing the results for the same query for the OR user model from Figure 2. See Figure 4. The different user model leads all weighted utilities to change, causing a change in the ordering of the overall set of options. In Figure 3, the highest ranked restaurant was Garden Cafe, mainly because of its good food quality (the attribute most highly valued by user CK). In contrast in Figure 4, the highest ranked restaurant is Junnos because of its reasonable cost, cost being the most highly valued attribute for user OR.

As mentioned above, WIT also allows the Wizard to straightforwardly change the query to add or remove constraints. Figure 5 shows the results of modifying the original query for user OR so that the price can now be a maxi-

mum of 70 dollars, and the food type can also be Seafood. Note that many more options are now made available.

#### 4. Experimental Method

There were four parts to the study: (a) generating representative domain tasks (b) acquiring user models (c) designing dialogue strategies that the Wizard could perform in real time that made good use of the user model; and (d) collecting sample dialogues to testing the utility of the user model/dialogue strategy combination.

Our procedure was as follows. We told users that they would be interacting with a Wizard (referred to as “Monsieur Croque”), who would simulate the functionality and strategies of the real dialogue system. We told them that the Wizard had access to information about thousands of restaurants in the New York area, derived from Zagats reviews including the following types of information: food type, food ratings, locations, prices, service, decor, along



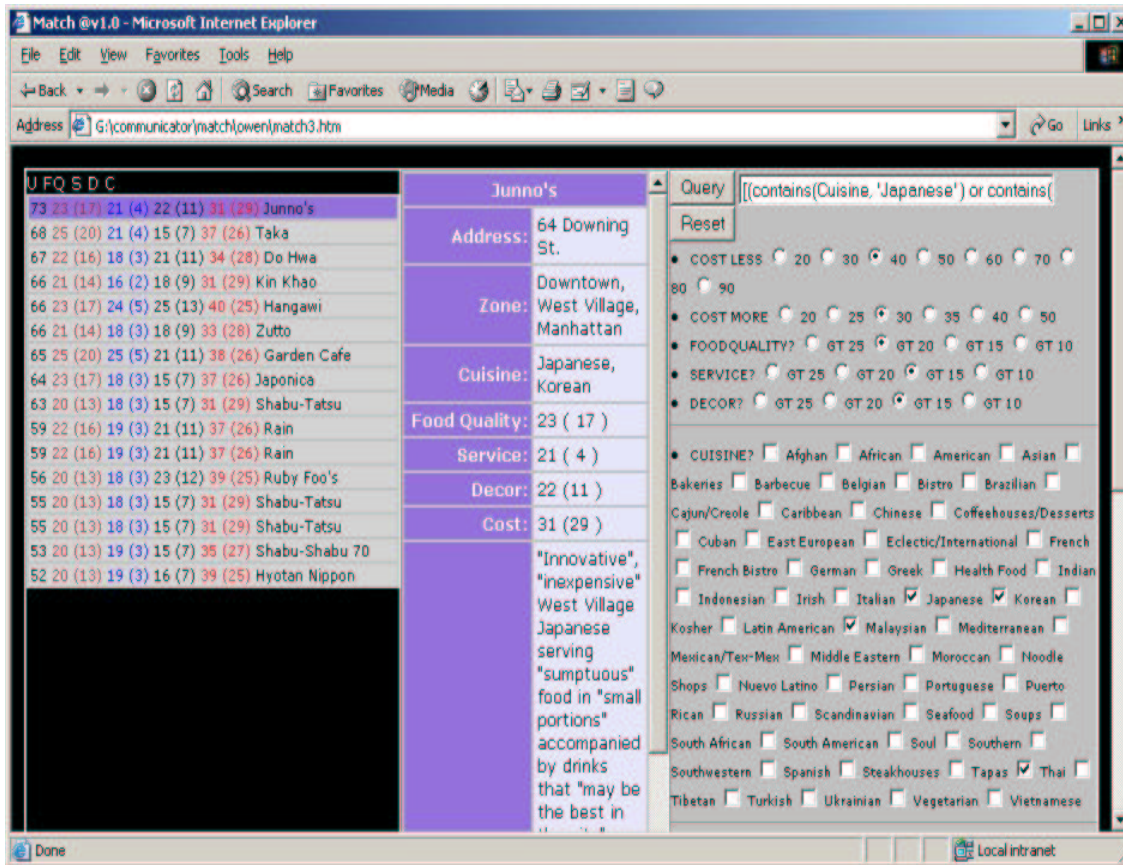


Figure 4: Wizard Interface for user OR after Wizard enters query for: *Japanese, Korean, Malaysian or Thai restaurants, costing between 30-40 dollars, with food quality greater than 20 and service and decor greater than 15 anywhere in Manhattan.*

with restaurant reviews that are composites of comments made by different Zagat survey participants.

#### 4.1. Generating Representative Tasks and Acquiring User Models

We gave 15 users an illustrative example of the information available for all restaurants and asked them to generate two sample task scenarios, according to the following description: *A scenario should be a description of a set of characteristics that would help Mr. Croque find a small set of restaurants that would match your description. Our goal is to examine the process by which you and Mr. Croque jointly identify a restaurant that you want to eat at so please do not select a particular restaurant in advance.* The initial instructions and scenario generation were carried out in email. Fifteen users responded with sample task scenarios. Two such tasks are in Figure 6; a sample dialogue for the CK task is in Figure 8.

#### 4.2. Wizard Dialogue Strategy

We now describe the strategies investigated in our experiment. The goal of a SUMMARY strategy is to provide an overview of the range of overall utility of the option set, along with the dimensions along which that set differ with respect to their attribute values. The aim is to inform users about both the range of choices, along with the range of reasons for making those choices. After entering a query corresponding to the user's choice in WIT, the Wizard examines the user selected set of restaurants and determines which attributes have the *same* values and which attributes have *different* values. Then he states the ways in which the restaurants are similar or different. The RECOMMENDATION then describes the first restaurant including all attributes that have not been mentioned so far. The tailored strategies are applied with the relevant user model.

The Zagat attributes are on a scale of 1-30. The Wiz-

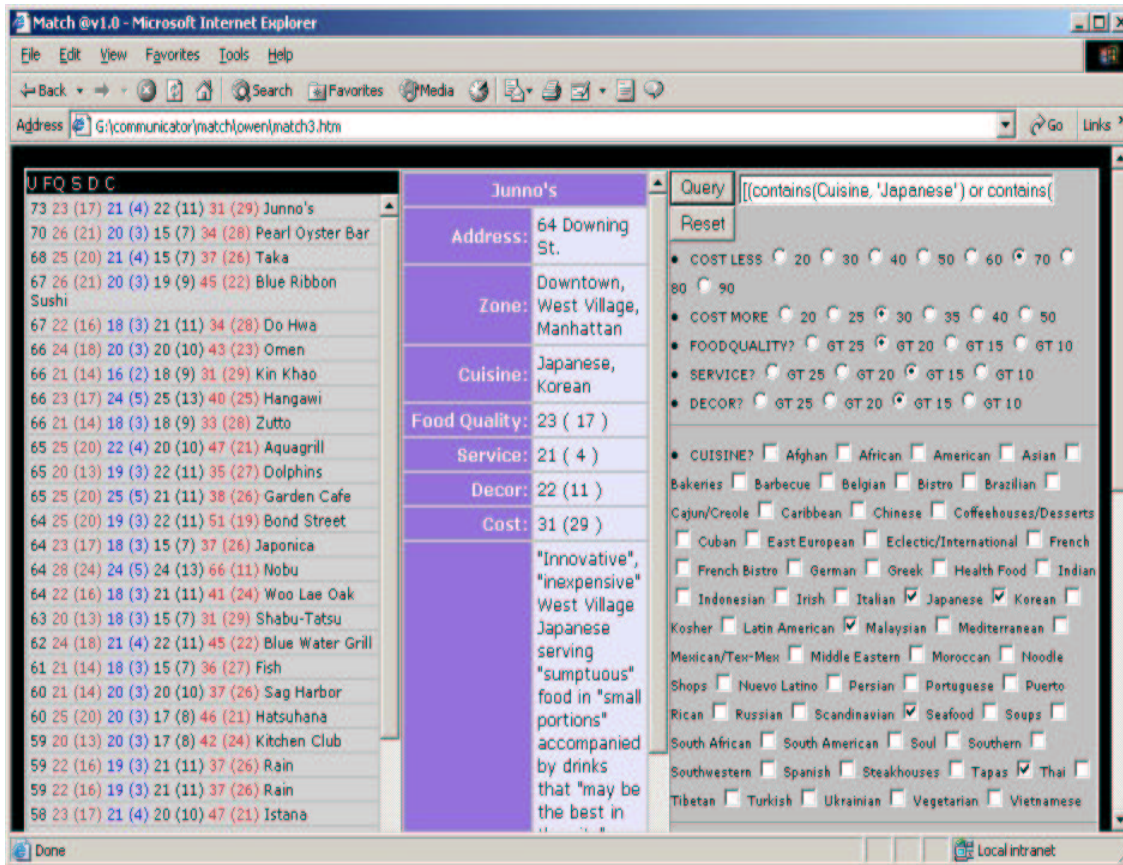


Figure 5: Wizard Interface for expanded query of cost up to \$70 and Seafood foodtype for the OR user

ard's dialogue strategy lexicalizes the absolute values in order to increase comprehensibility. We lexicalised them as follows: 26-30 excellent; 21-25 very good; 16-20 decent; 11-15 poor. We did not lexicalise price instead using absolute value, as there was little agreement about how to describe cost among pilot subjects. Two restaurants were judged to have the same value for a given attribute if the attribute had the same lexicalisation. We did not make similarity judgments about price.

Here is an example of a SUMMARY followed by a RECOMMEND: *There are 20 (if more than 20, say "many") restaurants that satisfy your criteria. The first three have decent decor, but differ in food quality, service and cost. The first one I have is the Garden Cafe, which is in midtown east. It's Japanese, it has very good food and service and the cost is 38 dollars.*

We contrasted this with the SERIAL strategy applied with a default user model derived by combining the average weights for the 15 user models we collected. The SERIAL strategy specified the number of restaurants satisfying

the query, and then stated the attributes in sequence, stating positive before negative values and aggregating across these where possible: *There are 18 restaurants that satisfy your criteria, the first one is Nyona, which is in Chinatown, it's southeast asian, the food quality is very good, although the decor and service are poor. The cost is 21 dollars.*

### 4.3. Collecting Sample Dialogues

Six subjects participated in the Wizard dialogue collection experiment resulting in a corpus of 24 dialogues. All of the subjects were familiar with Manhattan restaurants. We first examined the 30 typical tasks generated by our users. By identifying the common characteristics of these user-generated tasks, we generated two further control tasks for the domain. Each user participated in four tasks, two that they had generated themselves and two control tasks. We used this combination of user-generated and control tasks to combine ecological validity while controlling for task variability. User-generated tasks have the advantage of being both real and motivating, i.e. they are problems that the

| USER | TASK   |
|------|--|
| MS   | We want to go to the Indian restaurant with the best cuisine and the best service in walking distance of the Broadway theater district. We can't eat before 6, and we need to be able to leave the restaurant by 7:30 to make an 8 p.m. show near Times Square. Don and I will both arrive separately via subway, so transportation isn't an issue. We're willing to pay up to \$50 each for the meal, including drinks.   |
| CK   | I'm going to see the play Chicago on May 19. It is at the Shubert Theatre. I'm going to the matinee. Since this is a birthday celebration, we want to go out for an early dinner afterwards. I think a French restaurant in walking distance from there would be nice. My friends are wine experts, so it would be good if there was an impressive wine selection. I'm not too worried about the price, but I don't want to have to mortgage my house for this meal. |

Figure 6: Two Sample Tasks from Users MS and CK

user genuinely wants to solve. At the same time, however there was a great deal of variability in the complexity and number of solutions to these user-generated tasks, and we wanted to be able to control for this.

The underlying model and Wizard strategy were also varied; each user carried out two tasks with their own user model, and the tailored SUMMARY and RECOMMEND dialogue strategies. Each user also carried out two tasks with the default user model and the SERIAL strategy. Model/strategy and task provenance were crossed so that each user overall received four tasks: self-task/own model/tailored strategy, self task/default model/serial strategy, control task/own model/tailored strategy, control task/default model/serial strategy. Users carried out the four tasks in two separate sessions. Task order was randomised but each session included one user-generated and one control task, one own model/tailored strategy and one default model/serial strategy.

A sample dialogue illustrating a control task of *Find a restaurant in the West Village* using a strategy tailored to user HA is in Figure 7. The user model for user HA is in Figure 2. This dialogue illustrates the use of the user model and the requirements for the Wizard interface. For example, Figure 2 shows that user HA typically likes Japanese food, thus the most highly ranked restaurant first mentioned to this user is a Japanese restaurant. However, the user specifies a situational bias against Japanese because he has eaten Japanese food recently. Later in the dialogue he says that he has been eating too much Italian food lately. WIT allows the Wizard to specify what type of cuisine is **not** desired as

|      |   |
|------|---|
| Wiz: | I've actually got, erm, a large number of restaurants, again this time about, 50 I'd say, erm   |
| HA:  | do you know the location of the Lucille Lortel theatre? I should know I've just been at a play there.   |
| Wiz: | its on, erm, its between 6 and 7th avenues on, is it Nicholas? Christopher.   |
| HA:  | oh its, I see yes. Right yes, I know where it is.   |
| Wiz: | erm, so its west village, right, west of west village.  |
| HA:  | right   |
| Wiz: | so I have erm, the top three restaurants are erm again very similar in dcor, they have different food quality service and cost. And the first one I have is er a Japanese in the west village, erm, very good food and service, erm, the cost is \$37.  |
| HA:  | ok, since I just had Japanese   |
| Wiz: | you don't want Japanese?  |
| HA:  | I think I'll want something different today.  |
| Wiz: | ok. erm, so I've now got about 40 erm, the top three on this occasion, erm, all have very similar decor, but they have different food quality, service and cost, erm, the top one is erm, Mexicana Mama, erm, which is west village Mexican/Tex/mex which has er very good food, er poor decor, sorry very good service, and erm the cost is \$26. The next one is the Pearl Oyster Bar which is again west village, it's a seafood restaurant excellent quality food, erm, decent service, cost is \$34. |
| HA:  | can you read the review for that - does it say anything about how fresh the seafood is?   |
| Wiz: | its just a marble counter with a few small tables, but Pearl has won over its neighbourhood with its casual charm, and Maine inspired seafood.  |
| HA:  | hmm that's Maine as in the state?   |
| Wiz: | yes. erm the next one is erm, called the Blue Ribbon, erm this is again west village, it's a new American west with er very good food and service, the cost is \$45 a head. Do you want to hear some more?  |
| HA:  | well, ok, I actually I think that I like the erm, the Oyster,   |
| Wiz: | ok, the Pearl Oyster  |

Figure 7: Excerpt from the Wizard Dialogue with user HA, Tailored Strategy, for the West Village task

well as what types are desired in order to be able to quickly modify the query for situations such as these.

A dialogue illustrating the CK task in Figure 6 with the default user model and the SERIAL dialogue strategy is in Figure 8. This dialogue illustrates issues with what information the Wizard had available and the user's understanding of the system's capabilities. The user is trying to find a French restaurant for her friends who are food snobs. She would like to hear about the menu and wine list but this information is not available. The Wizard offers that he does



|      |   |
|------|---|
| Wiz: | So there are approximately 20 restaurants that satisfy the criteria that you specified. The first one is called Lespinasse and it's a French restaurant in midtown east which has excellent food dcor and service and the cost is \$86. |
| CK:  | do you have any sample menus or anything?   |
| Wiz: | I am sorry, I have no sample menus.   |
| CK:  | are they known for their wine list?   |
| Wiz: | again I am sorry I haven't got that information.  |
| CK:  | ok can you tell me about another restaurant then?   |
| Wiz: | I do have, I do have reviews.   |
| CK:  | for that Lespinasse restaurant.   |
| Wiz: | yes.  |
| CK:  | can you read me the review or part of it?   |
| Wiz: | The much-celebrated Gray Kunz has departed, leaving the more traditional Christian Louvrier at his place behind the stove.  |
| Wiz: | that's it.  |
| CK:  | (laughs) ok. Tell me about another restaurant   |
| Wiz: | ok the next restaurant is called La Grenouille, it's again a French restaurant. Again it's in midtown east, the food quality decor and service are all excellent the cost of this one is \$80.  |
| CK:  | ok do you have menus or reviews for this one.   |
| Wiz: | the review says: Gorgeous flowers, fine service, rich people and a menu written entirely in French.   |
| CK:  | I think that you need better better reviews. They don't. ok the food was excellent in both those places. Can you tell me about can you tell me how far those are from the Shubert theatre?  |
| Wiz: | That's 8 blocks.  |
| CK:  | 8 blocks so that's great. Ok. Do you have anything that's tell me about the best restaurant that you have.  |
| Wiz: | the best restaurant along what dimension?   |
| CK:  | erm   |
| Wiz: | in terms of food quality, cost..  |
| CK:  | food quality  |
| Wiz: | ok, erm the best quality restaurant is er Lespinasse  |
| CK:  | ok Lespinasse, that's the first one that you told me about.   |
| Wiz: | yes   |
| CK:  | ok, erm ,then I'm happy with that one.  |

Figure 8: Excerpt form the Wizard Dialogue with user CK, Default Strategy, CK Task from Figure 6

have reviews, but a little later, she says that better reviews are needed. The dialogue also illustrates how the Wizard needed access to distance information. The real MATCH system can do such calculations, but this was not implemented in WIT. The Wizard kept a map of New York City next to him during the dialogue interactions, and tried to quickly make such calculations.

For each dialogue we collected both quantitative and qualitative data. We collected quantitative information about the number of turns, words and duration of each dialogue. After each dialogue was completed, the users were asked to complete a survey. The survey first requested the users to give permission for their dialogues to become part of a public corpus so they can be distributed as part of the ISLE project. Then they were required to state their degree of agreement on a 5 point Likert scale with three specific statements designed to probe their perception of their interaction with the Wizard (Mr. Crocque): (1) I feel confident that I selected a good restaurant in this conversation; (2) Mr. Crocque made it easy to find a restaurant that I wanted to go to; and (3) I'd like to call Mr. Crocque regularly for restaurant information.

## 5. Results

### 5.1. Quantitative Results

Our results were as follows. First, we looked at the effects of the User Model/Dialogue Strategy on objective measures such as dialogue length and time to solution. Having the User Model/Dialogue Strategy did not affect the number of turns (respective means for User Model/Dialogue Strategy and Default Model/ Serial Strategy were 41 and 49, t test, ns) or words (respective means 478 and 461) in the dialogue, and there were also no differences in time to solution (7.7 mins and 6.3 minutes respectively). It was obvious however that there were huge amounts of variance in these measures, due to the fact that user tasks had very different complexity. Part of this was due to the nature of the scenarios chosen. Some user tasks had no solution in the database (e.g. high quality food, for a very low price), others had many solutions (Italian food in Greenwich Village). This led to extremely long dialogues for different reasons, either to modify the original scenario, so that at least one solution was possible, or to constrain the query to reduce the size of the original huge set. Because of the large variance in user generated tasks, we conducted a second analysis which included only control tasks. Again however, there were no differences resulting from the User Model/Dialogue Strategy.

There was, however, one difference following from providing the User Model/Dialogue Strategy. The User Model led users to take a more active role in negotiating solutions. The User Model led them to offer more constraints to the Wizard about their desired solution (respective means for User Model/Dialogue Strategy and Default Model/ Serial Strategy were 2.1 and 1.2, t test, significant).

## 5.2. Qualitative Results

We also looked at the effects of the User Model/Tailored Strategy on people's *perceptions of the system*. Again there were few differences, in (a) their confidence that they had selected a good restaurant (respective means for User Model/Tailored Strategy and Default Model/ Serial Strategy were 3.9 and 3.9, t test, ns), (b) their belief that the Wizard made it easy to find a good restaurant or (respective means for User Model/Tailored Strategy and Default Model/ Serial Strategy were 3.7 and 3.9, t test, ns) (c) that they would use the system regularly for restaurant information (respective means for User Model/Tailored Strategy and Default Model/ Serial Strategy were 3.8 and 3.2, t test, ns). Again we conducted a second analysis exclusively for control tasks alone, but found no differences.

Despite the failure of our strategy manipulation, there were some significant relationships between objective and subjective measures. Users who took longer to complete their task thought that tasks were harder ( $R=0.48$ ,  $p=0.02$ ), and people who had more verbose dialogues were less likely to think that they would want to use the system again ( $R=0.39$ ,  $p<0.05$ ).

Users also made a number of comments about the strategies. Our strategies were exhaustive in that they mentioned every attribute in the database. One user commented that he wanted a more terse summary: one that only included attributes that were important to him. For example, since he did not care about decor, there was little point in providing him information about this. Two users were unsure about how interactive they could be. They commented that they did not know whether they could interrupt the Wizard to ask further questions or to add constraints. Another user also wanted the system to make statements about metadata, informing her about what the system knew about so that she could constrain her questions to these topics. One user also wanted the system to retain a history of prior interactions, such as remembering the fact that she had been there before. Several users also felt that output could have been clustered in ways that would have made it easier to comprehend, such as by classifying sets of viable restaurants in terms of food type.

Finally there were a number of comments about the coverage that the system offered. One user noted that the system did not contain information about specific restaurants that she knew about and liked, and that this undermined her belief in the "system". Another user commented that there was other useful information, that she wanted to know about the restaurant, such as whether or not it took reservations or what the wine list was like (See Figure 8). Others wanted directions to restaurants from their current location. Several users placed great emphasis on the restaurant re-

views, and commented that the reviews were too short to allow reasoned decisions to be made. However as implementers we had no control over the information that was available as this was extracted automatically from available online data. We have, however, added support for directions and route finding in our working MATCH system.

## 6. Conclusions and Future Work

In conclusion, we have collected 30 representative tasks and 24 dialogues for a complex information seeking domain. We also have user models for 15 users in that domain. These resources will be made available as part of the ISLE project. We also devised a useful tool for supporting Wizard of Oz style data collection, that embodies user's specific preferences. This should support the collection of further data in this, and with suitable modifications, other domains.

Our main experimental manipulation was not completely successful: several of our predicted effects were not supported by our data. Nevertheless we did observe some interesting findings: Users were more proactive with the Own Model/Tailored combination, and we also found correlations between task length and task complexity, as well as between verbosity and likelihood of future use. The qualitative comments have also been useful for the further development of both the MATCH system and the strategies implemented in it.

There are several possible reasons for the lack of observed effects which we are investigating in current work. First we only ran a small number of subjects. Second, given the small number of subjects, the default user model may have been too close to some subject's models. In our current study we are manipulating the distance between user models in order to determine how similar two models have to be to affect user's perceptions of the system's response strategy. Finally, there are many tailored strategies that can be implemented. Here, we were constrained to experiment with strategies that the Wizard could produce in real time given the current version of WIT. We believe that a future version of WIT could do more user model calculations and provide more help to the Wizard, enabling us to experiment with more complex tailored strategies. In current work, we are running an automated, non-interactive version of the experiment comparing multiple strategies (compare, summarise, recommend) for text and speech presentation for a larger number of users.

## 7. Acknowledgements

The work reported in this paper was partially funded by DARPA contract MDA972-99-3-0003 and by the National

Science Foundation under Grant No. 9910603 (ISLE) to the University of Pennsylvania.

## 8. References

- Giuseppe Carenini and Johanna D. Moore. 2000. An empirical study of the influence of argument conciseness on argument effectiveness. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-00)*.
- Giuseppe Carenini and Johanna D. Moore. 2001. An empirical study of the influence of user tailoring on evaluative argument effectiveness. In *IJCAI*, pages 1307–1314.
- Ward Edwards and F. Hutton Barron. 1994. Smarts and smarter: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes*, 60:306–325.
- Michael Johnston, Srinivas Bangalore, and Gunaranjan Vasireddy. 2001. Match: Multimodal access to city help. In *Automatic Speech Recognition and Understanding Workshop*, Madonna Di Campiglio, Trento, Italy.
- R. Keeney and H. Raiffa. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons.
- Esther Levin, S Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. DiFabrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker. 2000. The at&t darpa communicator mixed-initiative spoken dialog system. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP00*.
- Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In *Proceedings of the Meeting of the Association of Computational Linguistics, ACL 2001*.